Bayesian Inference

Frank Schorfheide
Professor of Economics, University of Pennsylvania

September 14, 2021

Bayesian Inference

- Ingredients of Bayesian Analysis:
  - Likelihood function $p(Y|\theta)$
  - Prior density $p(\theta)$
  - Marginal data density $p(Y) = \int p(Y|\theta)p(\theta)d\theta$
- Bayes Theorem:

$$p(\theta|Y) = \frac{p(Y|\theta)p(\theta)}{p(Y)} \propto p(Y|\theta)p(\theta)$$

- Implementation: usually by generating a sequence of draws (not necessarily iid) from posterior

$$\theta^i \sim p(\theta|Y), \quad i = 1, \ldots, N$$

- Algorithms: direct sampling, accept/reject sampling, importance sampling, Markov chain Monte Carlo sampling, sequential Monte Carlo sampling...

Bayesian Inference for DSGE Models

- We previously discussed the evaluation of the likelihood function: given a parameter θ
  - solve the DSGE model to obtain the state-space representation;
  - use the Kalman filter to evaluate the likelihood function.
- Let's talk a bit about prior distributions.

Prior Distributions

- Ideally: probabilistic representation of our knowledge/beliefs before observing sample Y.
- More realistically: choice of prior as well as model are influenced by some observations. Try to keep influence small or adjust measures of uncertainty.
- Views about role of priors:
  1. keep them "uninformative" (???) so that posterior inherits shape of likelihood function;
  2. use them to regularize the likelihood function;
  3. incorporate information from sources other than Y;

Prior Elicitation for DSGE Models

- Group parameters:
  - steady-state related parameters
  - parameters assoc with exogenous shocks
  - parameters assoc with internal propagation
- Non-sample information $p(\theta|\mathcal{X}^0)$:
  - pre-sample information
  - micro-level information
- To guide the prior for θ, you can ask: what are its implications for observables Y?

Prior Distribution

Name	Domain	Prior		
		Density	Para (1)	Para (2)
Steady-State-Related Parameters $\theta_{(ss)}$				
$100(1/\beta-1)$	$\mathbb{R}^+$	Gamma	0.50	0.50
$100\log\pi^*$	$\mathbb{R}^+$	Gamma	1.00	0.50
$100\log\gamma$	$\mathbb{R}$	Normal	0.75	0.50
λ	$\mathbb{R}^+$	Gamma	0.20	0.20
Endogenous Propagation Parameters $\theta_{(endo)}$				
ζ_p	$[0,1]$	Beta	0.70	0.15
$1/(1+\nu)$	$\mathbb{R}^+$	Gamma	1.50	0.75

Notes: Marginal prior distributions for each DSGE model parameter. Para (1) and Para (2) list the means and the standard deviations for Beta, Gamma, and Normal distributions; the upper and lower bound of the support for the Uniform distribution; s and ν for the Inverse Gamma distribution, where $p_{\mathcal{IG}}(\sigma|\nu,s) \propto \sigma^{-\nu-1}e^{-\nu s^2/2\sigma^2}$. The joint prior distribution of θ is truncated at the boundary of the determinacy region.

Prior Distribution

Name	Domain	Prior		
		Density	Para (1)	Para (2)
		Exogenous Shock Parameters $\theta_{(exo)}$		
ρ_ϕ	$[0,1)$	Uniform	0.00	1.00
ρ_λ	$[0,1)$	Uniform	0.00	1.00
ρ_z	$[0,1)$	Uniform	0.00	1.00
$100\sigma_\phi$	$\mathbb{R}^+$	InvGamma	2.00	4.00
$100\sigma_\lambda$	$\mathbb{R}^+$	InvGamma	0.50	4.00
$100\sigma_z$	$\mathbb{R}^+$	InvGamma	2.00	4.00
$100\sigma_r$	$\mathbb{R}^+$	InvGamma	0.50	4.00

Notes: Marginal prior distributions for each DSGE model parameter. Para (1) and Para (2) list the means and the standard deviations for Beta, Gamma, and Normal distributions; the upper and lower bound of the support for the Uniform distribution; s and ν for the Inverse Gamma distribution, where $p_{\mathcal{IG}}(\sigma|\nu,s) \propto \sigma^{-\nu-1}e^{-\nu s^2/2\sigma^2}$. The joint prior distribution of θ is truncated at the boundary of the determinacy region.

Draws from Posterior

- We will focus on Markov chain Monte Carlo (MCMC) algorithms that generate draws $\{\theta^i\}_{i=1}^N$ from posterior distributions of parameters.

- Draws can then be transformed into objects of interest, $h(\theta^i)$, and under suitable conditions a Monte Carlo average of the form

$$\bar{h}_N = \frac{1}{N}\sum_{i=1}^{N} h(\theta^i) \approx \mathbb{E}_\pi[h] = \int h(\theta)p(\theta|Y)d\theta.$$

- Strong law of large numbers (SLLN), central limit theorem (CLT)...

Markov Chain Monte Carlo (MCMC)

- Main idea: create a sequence of serially correlated draws such that the distribution of θ^i converges to the posterior distribution $p(\theta|Y)$.
- Some Intuition: suppose we generate draws from the process

$$\theta^i = \rho\theta^{i-1} + \sqrt{1-\rho^2}\epsilon^i, \quad \epsilon^i \sim N(0,1), \quad \theta^0 = 0.$$

Then,

- The θ^i draws are serially correlated.
- Provided $|\rho| < 1$, the effect of the initialization $\theta^0 = 0$ will die out eventually, and $\theta^i \approx N(0,1)$.
- $\frac{1}{N}\sum_{i=1}^N \theta^i \xrightarrow{p} \mathbb{E}[\theta] = 0$.
- The closer ρ to zero, the more accurate the Monte Carlo approximation.

Generic Metropolis-Hastings Algorithm

For $i = 1$ to N:

1. Draw ϑ from a density $q(\vartheta|\theta^{i-1})$.
2. Set $\theta^i = \vartheta$ with probability

$$\alpha(\vartheta|\theta^{i-1}) = \min\left\{1, \ \frac{p(Y|\vartheta)p(\vartheta)/q(\vartheta|\theta^{i-1})}{p(Y|\theta^{i-1})p(\theta^{i-1})/q(\theta^{i-1}|\vartheta)}\right\}$$

and $\theta^i = \theta^{i-1}$ otherwise.

Note that

$$\frac{p(\vartheta|Y)}{p(\theta|Y)} = \frac{p(Y|\vartheta)p(\vartheta)/p(Y)}{p(Y|\theta)p(\theta)/p(Y)} = \frac{p(Y|\vartheta)p(\vartheta)}{p(Y|\theta)p(\theta)}$$

We draw θ^i conditional on a parameter draw θ^{i-1}: leads to Markov transition kernel $K(\theta|\tilde{\theta})$.

Benchmark Random-Walk Metropolis-Hastings (RWMH) Algorithm for DSGE Models

- Initialization:
  1. Use a numerical optimization routine to maximize the log posterior, which up to a constant is given by $\ln p(Y|\theta) + \ln p(\theta)$. Denote the posterior mode by $\hat{\theta}$.
  2. Let $\hat{\Sigma}$ be the inverse of the (negative) Hessian computed at the posterior mode $\hat{\theta}$, which can be computed numerically.
  3. Draw θ^0 from $N(\hat{\theta}, c_0^2\hat{\Sigma})$ or directly specify a starting value.

- Main Algorithm – For $i = 1, \ldots, N$:
  1. Draw ϑ from the proposal distribution $N(\theta^{i-1}, c^2\hat{\Sigma})$.
  2. Set $\theta^i = \vartheta$ with probability

$$\alpha(\vartheta|\theta^{i-1}) = \min\left\{1, \frac{p(Y|\vartheta)p(\vartheta)}{p(Y|\theta^{i-1})p(\theta^{i-1})}\right\}$$

     and $\theta^i = \theta^{i-1}$ otherwise.

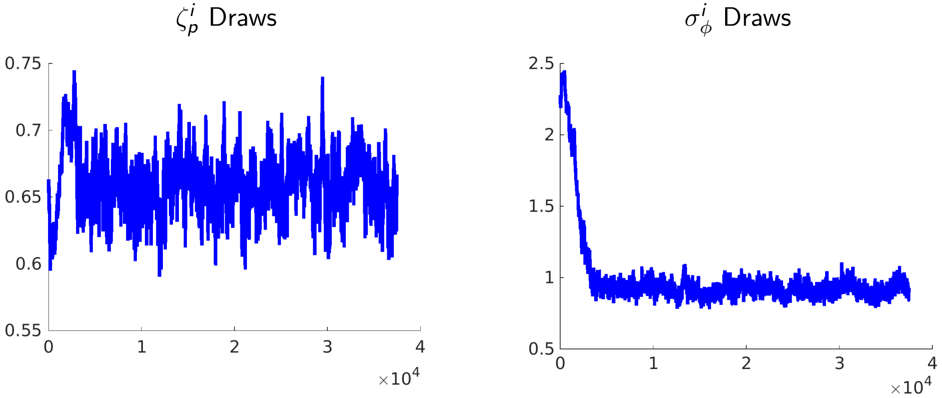
Benchmark RWMH Algorithm for DSGE Models

- Initialization steps can be modified as needed for particular application.
- If numerical optimization does not work well, one could let $\hat{\Sigma}$ be a diagonal matrix with prior variances on the diagonal.
- Or, $\hat{\Sigma}$ could be based on a preliminary run of a posterior sampler.
- It is good practice to run multiple chains based on different starting values.

Numerical Illustration

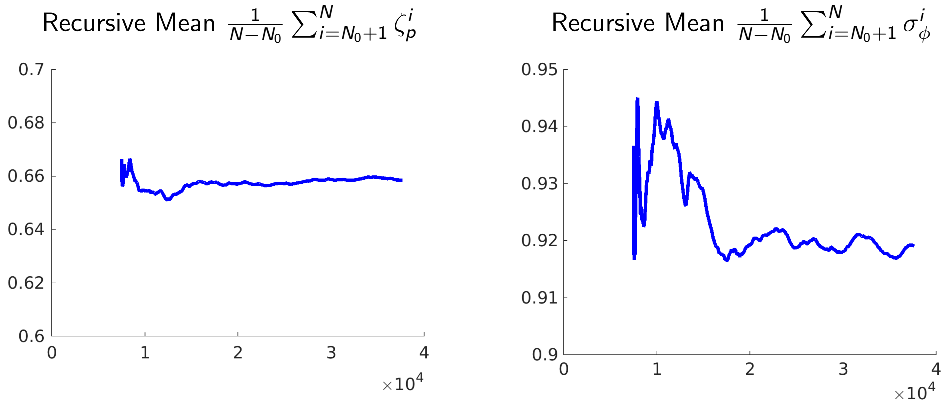
- Generate a single sample of size $T = 80$ from the stylized DSGE model.
- Combine likelihood and prior to form posterior.
- Draws from this posterior distribution are generated using the RWMH algorithm.
- Chain is initialized with a draw from the prior distribution.
- The covariance matrix $\hat{\Sigma}$ is based on the negative inverse Hessian at the mode. The scaling constant c is set equal to 0.075, which leads to an acceptance rate for proposed draws of 0.55.

Parameter Draws from MH Algorithm



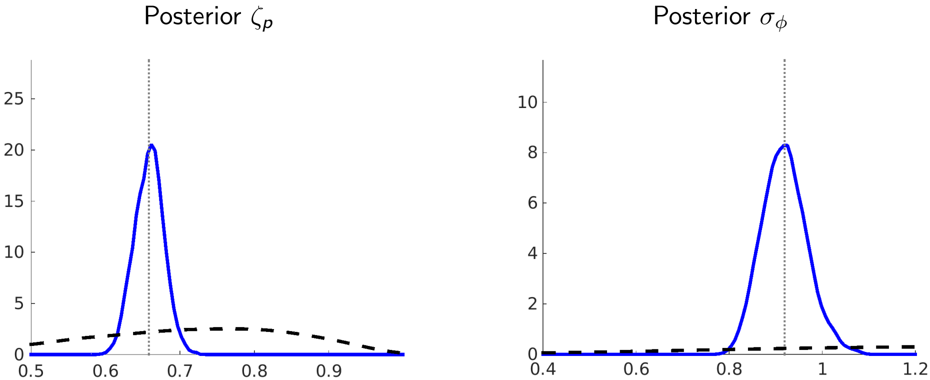
Notes: The posterior is based on a simulated sample of observations of size $T = 80$. The top panel shows the sequence of parameter draws and the bottom panel shows recursive means.

Parameter Draws from MH Algorithm



Notes: The posterior is based on a simulated sample of observations of size $T = 80$. The top panel shows the sequence of parameter draws and the bottom panel shows recursive means.

Prior and Posterior Densities



Notes: The dashed lines represent the prior densities, whereas the solid lines correspond to the posterior densities of ζ_p and σ_ϕ. The posterior is based on a simulated sample of observations of size $T = 80$. We generate $N = 37,500$ draws from the posterior and drop the first $N_0 = 7,500$ draws.

Why Does it Work?

- Algorithm generates a Markov transition kernel $K(\theta|\tilde{\theta})$: it takes a draw θ^{i-1} and uses some randomization to turn it into a draw θ^i.
- Important invariance property: if θ^{i-1} is from posterior $p(\theta|Y)$, then θ^i's distribution will also be $p(\theta|Y)$.
- Contraction property: if θ^{i-1} is from some distribution $\pi_{i-1}(\theta)$, then the discrepancy between the "true" posterior and

$$\pi_i(\theta) = \int K(\theta|\tilde{\theta})\pi_{i-1}(\tilde{\theta})d\tilde{\theta}$$

is smaller than the discrepancy between $\pi_{i-1}(\theta)$ and $p(\theta|Y)$.

Example: Convergence

- Define the Monte Carlo estimate

$$\bar{h}_N = \frac{1}{N}\sum_{i=1}^{N} h(\theta^i).$$

- Deduce from CLT

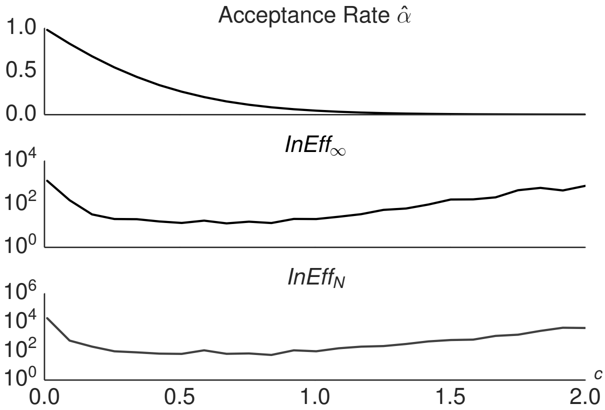
$$\sqrt{N}(\bar{h}_N - \mathbb{E}_\pi[h]) \Longrightarrow N\big(0, \Omega(h)\big), \quad \Omega(h) = \mathbb{V}_\pi[h] + \lim_{N \longrightarrow \infty} \frac{1}{N}\sum_{i=1}^{N}\sum_{j \neq i} COV\left[h(\theta^i), h(\theta^j)\right]$$

where $\Omega(h)$ is the long-run covariance matrix.

- In turn, the asymptotic inefficiency factor is given by

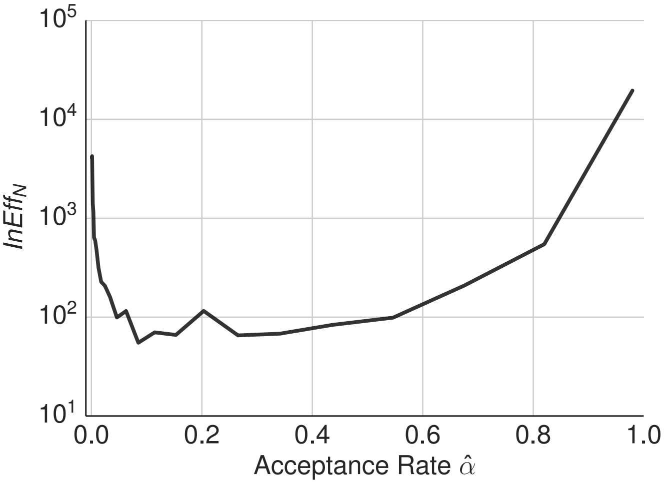
$$\text{InEff}_\infty = \frac{\Omega(h)}{\mathbb{V}_\pi[h]}.$$

DSGE Model Estimation: Effect of Scaling Constant c



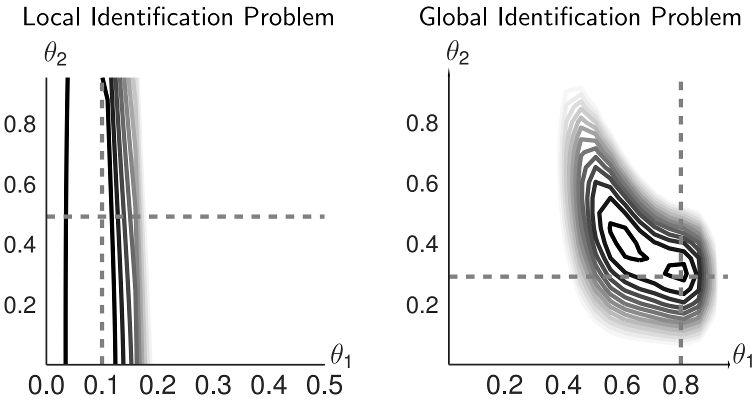
Notes: Results are based on $N_{run} = 50$ independent Markov chains. The acceptance rate (average across multiple chains), HAC-based estimate of $\text{InEff}_\infty[\bar{\tau}]$ (average across multiple chains), and $\text{InEff}_N[\bar{\tau}]$ are shown as a function of the scaling constant c.

DSGE Model Estimation: Acceptance Rate $\hat{\alpha}$ versus Inaccuracy InEff_N



Notes: $\text{InEff}_N[\bar{\tau}]$ versus the acceptance rate $\hat{\alpha}$.

Challenges Due to Irregular Posteriors

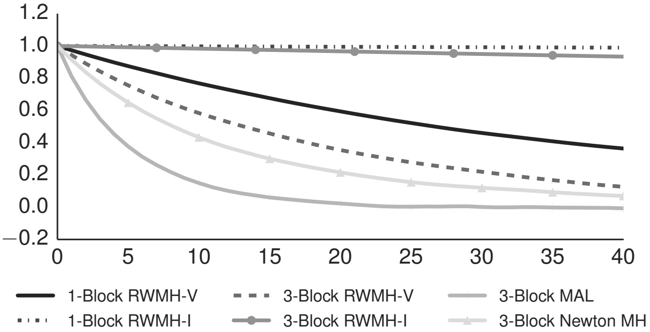


Notes: Intersections of the solid lines indicate parameter values that were used to generate the data from which the posteriors are constructed.

Improvements to MCMC: Blocking

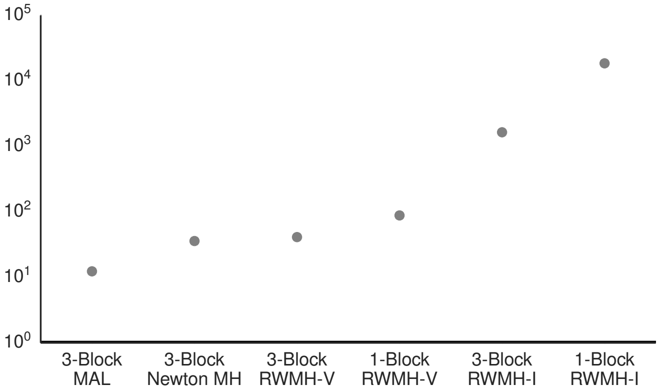
- In high-dimensional parameter spaces the RWMH algorithm generates highly persistent Markov chains which imply slow convergence of Monte Carlo averages (poor MCMC approximation).
- Potential Remedy:
  - Partition $\theta = [\theta_1, \ldots, \theta_K]$.
  - Iterate over conditional posteriors $p(\theta_k | Y, \theta_{<-k>})$.
- To reduce persistence of the chain, try to find partitions such that parameters are strongly correlated within blocks and weakly correlated across blocks or use random blocking.

Autocorrelation Function of τ^i



Notes: The autocorrelation functions are computed based on a single run of each algorithm.

Inefficiency Factor $\text{InEff}_N[\bar{\tau}]$



Notes: The small sample inefficiency factors are computed based on $N_{run} = 50$ independent runs of each algorithm.

Run Times and Tuning Constants for MH Algorithms

Algorithm	Run Time [hh:mm:ss]	Acceptance Rate	Tuning Constants
1-Block RWMH-I	00:01:13	0.28	$c = 0.015$
1-Block RWMH-V	00:01:13	0.37	$c = 0.400$
3-Block RWMH-I	00:03:38	0.40	$c = 0.070$
3-Block RWMH-V	00:03:36	0.43	$c = 1.200$

Notes: In each run we generate $N = 100,000$ draws. We report the fastest run time and the average acceptance rate across $N_{run} = 50$ independent Markov chains.

IID Equivalent Draws Per Second

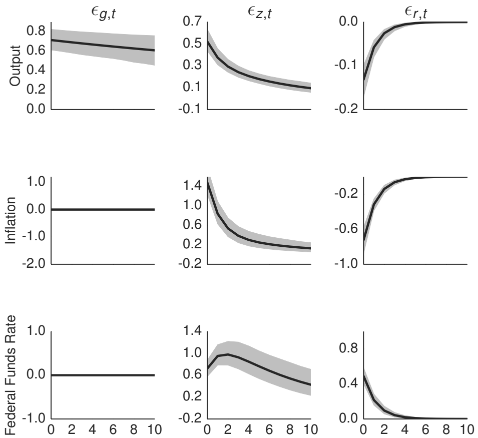
$$\textit{iid}\text{-equivalent draws per second} = \frac{N}{\text{Run Time [seconds]}} \cdot \frac{1}{\text{InEff}_N}.$$

- 1-Block RWMH-V: 7.76
- 3-Block RWMH-V: 5.65
- 3-Block RWMH-I: 0.14
- 1-Block RWMH-I: 0.04

What Can We Do With Our Posterior Draws?

- Store them on our harddrive!
- Convert them into objects of interest:
  - impulse response functions;
  - government spending multipliers;
  - welfare effects of target inflation rate changes;
  - forecasts;
  - (...)

Parameter Transformations: Impulse Responses



Notes: The figure depicts pointwise posterior means and 90% credible bands. The responses of output are in percent relative to the initial level, whereas the responses of inflation and interest rates are in annualized percentages.

Bayesian Inference – Decision Making

- The posterior expected loss of decision $\delta(\cdot)$:

$$\rho\big(\delta(\cdot)|Y\big) = \int_\Theta L\big(\theta, \delta(Y)\big) p(\theta|Y) d\theta.$$

- Bayes decision minimizes the posterior expected loss:

$$\delta^*(Y) = \text{argmin}_d \; \rho\big(\delta(\cdot)|Y\big).$$

- Approximate $\rho\big(\delta(\cdot)|Y\big)$ by a Monte Carlo average

$$\bar{\rho}_N\big(\delta(\cdot)|Y\big) = \frac{1}{N}\sum_{i=1}^{N} L\big(\theta^i, \delta(\cdot)\big).$$

- Then compute

$$\delta_N^*(Y) = \text{argmin}_d \; \bar{\rho}_N\big(\delta(\cdot)|Y\big).$$

Bayesian Inference

- Point estimation:
  - Quadratic loss: posterior mean
  - Absolute error loss: posterior median
- Interval/Set estimation $\mathbb{P}_\pi\{\theta \in C(Y)\} = 1 - \alpha$:
  - highest posterior density sets
  - equal-tail-probability intervals

Posterior Model Odds and Marginal Data Densities

- Posterior model probabilities can be computed as follows:

$$\pi_{i,T} = \frac{\pi_{i,0} p(Y|\mathcal{M}_i)}{\sum_j \pi_{j,0} p(Y|\mathcal{M}_j)}, \quad j = 1, \ldots, 2, \tag{1}$$

- where

$$p(Y|\mathcal{M}) = \int p(Y|\theta, \mathcal{M}) p(\theta|\mathcal{M}) d\theta \tag{2}$$

- Note:

$$\ln p(Y_{1:T}|\mathcal{M}) = \sum_{t=1}^{T} \ln \int p(y_t|\theta, Y_{1:t-1}, \mathcal{M}) p(\theta|Y_{1:t-1}, \mathcal{M}) d\theta$$

- Posterior odds and Bayes Factor

$$\frac{\pi_{1,T}}{\pi_{2,T}} = \underbrace{\frac{\pi_{1,0}}{\pi_{2,0}}}_{\text{Prior Odds}} \times \underbrace{\frac{p(Y|\mathcal{M}_1)}{p(Y|\mathcal{M}_2)}}_{\text{Bayes Factor}} \tag{3}$$

Computation of Marginal Data Densities: Modified Harmonic Mean

- Consider the following identity:

$$\frac{1}{p(Y)} = \int \frac{f(\theta)}{p(Y|\theta)p(\theta)} p(\theta|Y)d\theta,$$

  where $\int f(\theta)d\theta = 1$.

- Conditional on the choice of $f(\theta)$ an obvious estimator is

$$\hat{p}_G(Y) = \left[\frac{1}{N}\sum_{i=1}^{N} \frac{f(\theta^i)}{p(Y|\theta^i)p(\theta^i)}\right]^{-1},$$

  where θ^i is drawn from the posterior $p(\theta|Y)$.

- Geweke (1999):

$$\begin{aligned} f(\theta) &= \tau^{-1}(2\pi)^{-d/2}|V_\theta|^{-1/2}\exp\left[-0.5(\theta-\bar{\theta})'V_\theta^{-1}(\theta-\bar{\theta})\right] \\ &\quad \times \left\{(\theta-\bar{\theta})'V_\theta^{-1}(\theta-\bar{\theta}) \le F^{-1}_{\chi^2_d}(\tau)\right\}. \end{aligned}$$